ELSEVIER

Ecological Indicators 1 (2001) 49–60

ECOLOGICAL INDICATORS

www.elsevier.com/locate/ecolind

Strategies for evaluating indicators based on guidelines from the Environmental Protection Agency's Office of Research and Development

Janis C. Kurtz [a,*], Laura E. Jackson [b], William S. Fisher [a]

[a] *United States Environmental Protection Agency, Gulf Ecology Division, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Gulf Breeze, FL 32561, USA*

[b] *United States Environmental Protection Agency, Environmental Monitoring and Assessment Program, Office of Research and Development, National Health and Environmental Effects Research Laboratory, Research Triangle Park, NC 27711, USA*

Accepted 7 March 2001

Abstract

The Environmental Protection Agency's Office of Research and Development (ORD) has prepared technical guidelines to evaluate the suitability of ecological indicators for monitoring programs. The guidelines were adopted by ORD to provide a consistent framework for indicator review, comparison and selection, and to provide direction for research on indicator development. The guidelines were organized within four evaluation phases: (1) conceptual relevance; (2) feasibility of implementation; (3) response variability; (4) interpretation and utility. Three example indicators were analyzed to illustrate the use of the guidelines in an evaluation. The examples included a direct chemical measurement (dissolved oxygen concentration), an estuarine benthic community index, and a stream fish community index of biotic integrity. Comparison of the three examples revealed differences in approach, style and types of information used to address each guideline. The *Evaluation Guidelines* were intended to be flexible within a consistent framework and the various strategies used in the examples demonstrate that the process can be useful for a wide variety of indicators and program objectives. Published by Elsevier Science Ltd.

Keywords: Ecological indicator; Environmental condition; Environmental monitoring; Environmental assessment

1. Introduction

Escalating concern about environmental condition has motivated efforts to monitor and assess environmental status and trends. Early monitoring efforts focused on obvious, discrete sources of stress such as point source chemical emissions. It soon became evident that remote and combined stressors, while more difficult to measure, also significantly altered environmental condition. Since measuring and interpreting all factors and variables that interplay in ecological issues was impossible, monitoring programs began to develop and select indicators that measured characteristics of the most valued ecological components and those that were most responsive to a diversity of stressors. These ecological indicators included biological, chemical, or physical measurements, and indices or models that attempted to characterize or summarize critical and usually complex components of an ecosystem.

Indicators are signs or signals that relay a complex message, from potentially numerous sources, in a

* Corresponding author. Tel.: +1-850-934-9212; fax: +1-850-934-2402.
E-mail address: kurtz.jan@epa.gov (J.C. Kurtz).

1470-160X/01/$ – see front matter. Published by Elsevier Science Ltd.
PII: S1470-160X(01)00004-8

simple and useful manner. Due to the variety of environmental issues, the complexity of environmental data, and the necessity for management decisions, many types of indicators have been developed for many different purposes. They can reflect biological, chemical and physical aspects of ecological condition, and have been used to characterize status, track or predict change, identify stressors or stressed systems, assess risk, and influence management actions. Indicators have been used to describe past, present or future conditions at a variety of geographical scales and for a variety of valued resources. Because they are so diversified, development and selection of successful ecological indicators has become a relatively complex process.

The Environmental Protection Agency's Office of Research and Development (ORD) has prepared technical guidance (*Evaluation Guidelines*; US EPA, 2000) to assist with the development and selection of indicators for use in specific monitoring programs, particularly the EPA's Environmental Monitoring and Assessment Program (EMAP). This guidance was generated to improve indicator development and to facilitate indicator evaluation. The *Evaluation Guidelines* recommended that evaluation of an indicator include information and data supporting the indicator in the context of 15 guidelines. Researchers can use the guidelines informally to target gaps in knowledge and formulate future research directions. Managers can use the guidelines to select among existing indicators based on their resources and the objectives of their programs. While providing a consistent framework to address indicator issues, the guidelines are flexible enough to meet the needs of diverse environmental programs.

The 15 guidelines are organized into four phases (Table 1) that are functionally related (US EPA, 1994) and allow users to focus on four fundamental questions:

1. *Phase* 1: Conceptual relevance — Is the indicator relevant to the assessment question (management concern) and to the ecological resource or function at risk?
2. *Phase* 2: Feasibility of implementation — Are the methods for sampling and measuring the environmental variables technically feasible, appropriate, and efficient for use in a monitoring program?
3. *Phase* 3: Response variability — Are errors of measurement and natural variability over time and space sufficiently understood and documented?
4. *Phase* 4: Interpretation and utility — Will the indicator convey information on ecological condition that is meaningful to environmental decision-making?

Table 1
Overview of the evaluation guidelines for ecological indicators

Phase 1: Conceptual relevance
Guideline 1: Relevance to the assessment
Guideline 2: Relevance to ecological function
Phase 2: Feasibility of implementation
Guideline 3: Data collection methods
Guideline 4: Logistics
Guideline 5: Information management
Guideline 6: Quality assurance
Guideline 7: Monetary costs
Phase 3: Response variability
Guideline 8: Estimation of measurement error
Guideline 9: Temporal variability (within-season)
Guideline 10: Temporal variability (across-year)
Guideline 11: Spatial variability
Guideline 12: Discriminatory ability
Phase 4: Interpretation and utility
Guideline 13: Data quality objectives
Guideline 14: Assessment thresholds
Guideline 15: Linkage to management action

The phases progress from consideration of fundamental concepts to methodology, ability to distinguish differences over time or space, and application to program objectives. Movement from one phase to the next should highlight strengths or weaknesses of an indicator at its current state of development. Yet, in actual practice, application of the guidelines may be iterative rather than sequential. User discretion is reinforced through the decision to make individual guidelines neither essential nor optional. For some purposes, users may be willing to accept weaknesses in an indicator if it provides important information. Such a conclusion should be a conscious decision, not one based on a lack of information.

2. Example indicators

The *Evaluation Guidelines* applied three example indicators to the 15 guidelines to illustrate their

intended use. The indicators included (1) a direct chemical measurement (dissolved oxygen) to determine the spatial extent of hypoxia in estuarine waters, (2) a benthic index of estuarine condition that incorporated measures of community composition and diversity to distinguish areas of degraded and undegraded environmental conditions, and (3) a multimetric indicator of ecological condition based on stream fish assemblages. The examples varied in complexity, type of information, and extent of analysis provided for each guideline. Although the purpose of the *Evaluation Guidelines* was to provide a consistent evaluation framework, which is essential to comparative evaluation and indicator improvement, comparison of the three examples demonstrates the flexibility of the guidelines for different indicators and program objectives.

3. Strategies used in the indicator examples

From the data provided in support of the example indicators, different strategic approaches were taken to substantiate the tenents of individual guidelines. Seven selections from the 15 guidelines follow highlighting some of these differences.

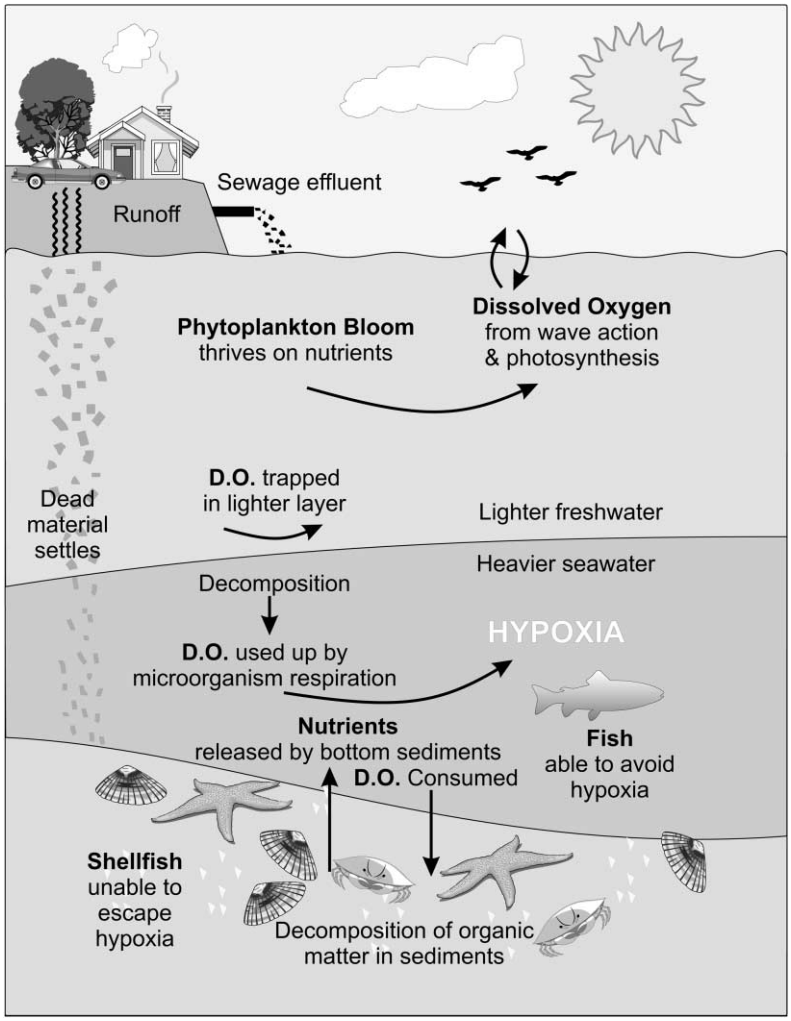

Fig. 1. Conceptual model showing the ecological relevance of dissolved oxygen concentration in estuarine water.

3.1. Guideline 1: Relevance to the assessment

More often than not, environmental research tracks mechanistic, rather than applied, scientific questions. This leads researchers to develop interesting concepts and measurements that might, or should, lead to useful indicators. Yet, these potential indicators have little chance for success unless they address valid and relevant assessment questions (Bardwell, 1991; Cowling, 1992; US EPA, 1994; Thornton et al., 1994). An assessment question is intended to address a critical aspect of a valued environmental resource. In the *Evaluation Guidelines*, each of the three example indicators were developed in response to an assessment question generated by EMAP: "what percent of estuarine area is hypoxic/anoxic?" (hypoxia indicator); "what percent of estuarine area has degraded benthic condition" (benthic condition index); and "what percent of stream miles have fish assemblages that differ from reference condition?" (fish assemblage indicator). The examples, though differing in approach and substance, were designed to provide quantitative information, used alone or in combination with other indicators, to optimize responsible management decisions.

3.2. Guideline 2: Relevance to ecological function

To properly interpret indicator results, it is imperative that there is a sound and defensible linkage between the indicator and the ecological function or critical resource it is intended to represent. For the hypoxia indicator, a conceptual model of oxygen dynamics in an estuarine ecosystem was illustrated graphically (Fig. 1). The illustration showed how hypoxic conditions form as oxygen enters an estuary from the atmosphere or as a product of phytoplankton photosynthesis, and becomes stratified and then depleted by bacterial decomposition. The benthic index example included an illustration of environmental stressors in estuaries and a diagram (Fig. 2) to show the linkages among sources of stress, types of stress and anticipated effects on community measures. A similar approach was used for the benthic index and fish assemblage examples, using both an illustration to identify the major structural and functional components of the ecosystem and a series of diagrams to characterize the anticipated effects of various stressors on the indicator metrics.

3.3. Guideline 3: Data collection methods

This guideline provided an opportunity to describe the methods of the indicator in some detail and point out differences with similar indicators described in the literature. It was also an opportunity to defend the selection of measurement parameters. Whereas relatively standard data collection methods were employed for the hypoxia indicator and the benthic index, the fish assemblage example used Guideline 3

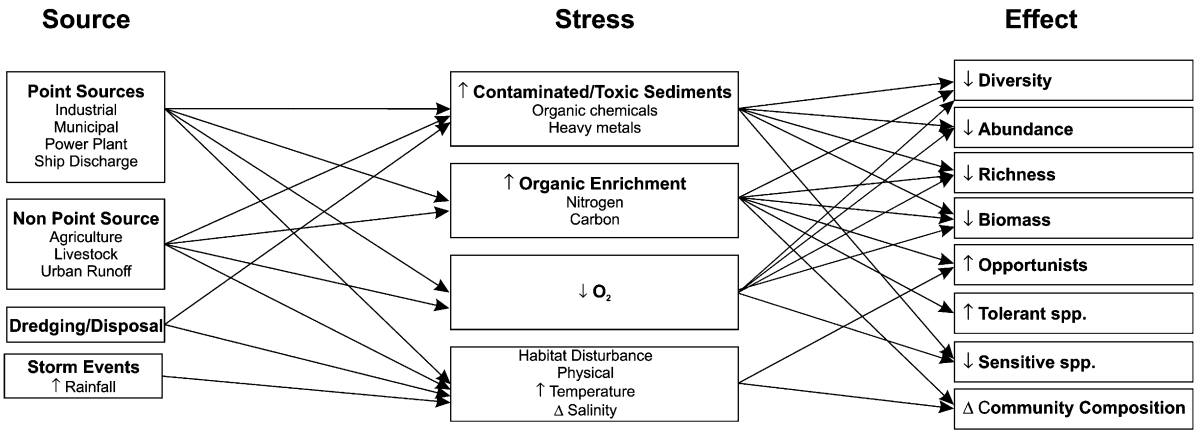

Fig. 2. Conceptual model of the benthic index, showing linkages between sources of stress, types of stress, and effects on the benthic community.

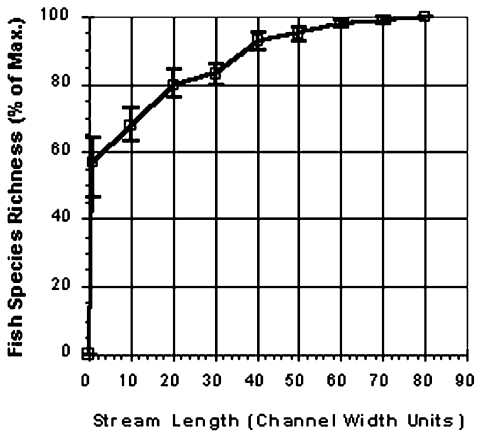

Fig. 3. The fish assemblage indicator used an effort–return curve of fish species richness vs. length of stream sampled to determine the size of the sample area.

to document the appropriate area (stream reach) to be sampled. A sample of fish were collected during a single pass through a prescribed length of stream (Karr et al., 1986; McCormick and Hughes, 1998). A pilot study showed that repeated sampling of a stream reach was neither practical nor representative, so it was important to determine that the length of stream sampled maximized the number of species collected. Based on preliminary sampling trials (Fig. 3), it was concluded that a stream length of 40 times the mean channel width was optimal.

3.4. Guideline 8: Estimation of measurement error

Measurement error is variability introduced through collecting, transporting and analyzing samples, and includes both human and instrument performance. Recognizing and characterizing measurement error is essential for determining whether or not an indicator has achieved data quality objectives of a program. In the hypoxia example, EMAP field crews performed a comparison to evaluate and, ultimately, minimize measurement error by measuring dissolved oxygen (DO) concentration with two different instruments. A frequency distribution of 784 stations (Fig. 4) showed the absolute difference between DO measurements collected by the more sophisticated CTD technology (instruments designed to measure conductivity, temperature and depth, and equipped with DO probes) versus measurements with a common DO meter. The data were collected over a 3-year period by nine different field crews. Consequently, the frequency distribution illustrates the total measurement error, that associated with the instrumentation as well as with operation of the instruments by different crews. Out of 784 stations, the data quality objective of $\leq$0.5 mg/l difference between instrument measurements was met over 90% of the time. No bias was detected, meaning the CTD values were neither consistently higher nor lower than the values reported from the DO meter. In this case, a second method of

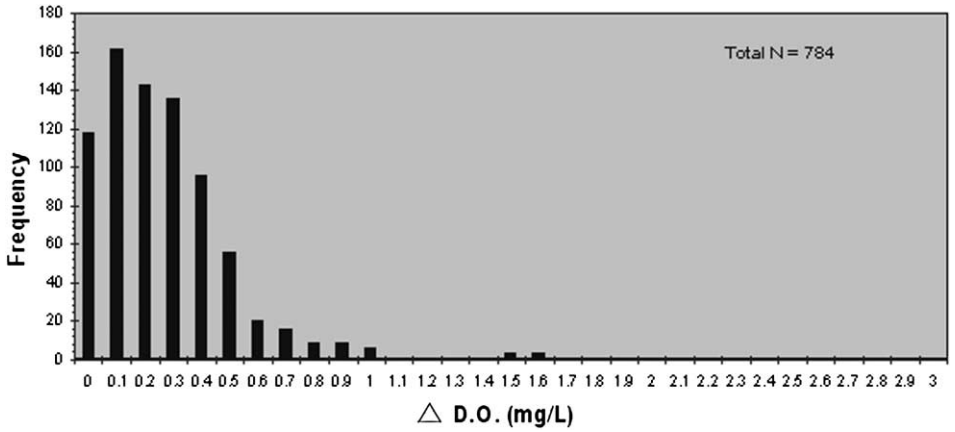

Fig. 4. Frequency distribution of dissolved oxygen (DO) at different EMAP sites (hypoxia indicator, Guideline 3). ΔDO represents the absolute difference between the CTD measurement and that from a second instrument. Over 90% of the stations met the measurement quality objective ($\Delta DO \leq 0.5$ mg/l).

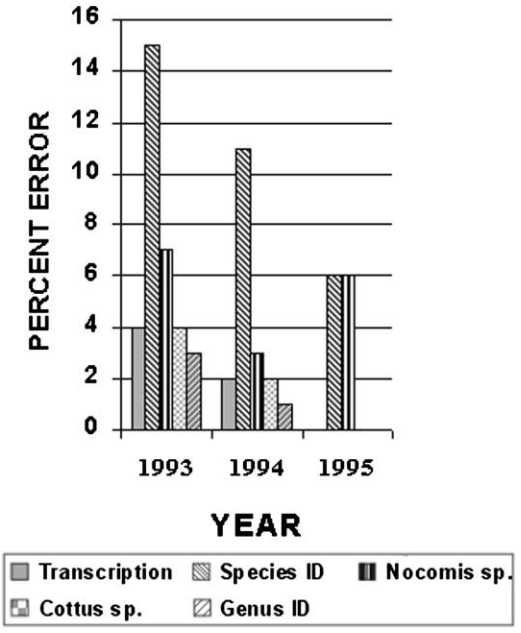

Fig. 5. Comparison of different types of measurement errors encountered in the fish assemblage indicator.

measurement was used to characterize measurement error.

In the fish assemblage example, incorrect identification of fish species was considered the most critical measurement error. Various means of controlling this source of error were considered, including the collection and confirmation of voucher specimens, the use of personnel experienced in fish identification, and additional training for field identification of regional fishes. An evaluation of measurement error was developed using data developed from 3 years of sampling, where five types of error were investigated (Fig. 5). Transcription errors occurred when the incorrect species code or common name was recorded, whereas four other types consisted of errors in taxonomic identification. Improvements were made over the 3-year period, including virtual elimination of transcription errors, misidentification of sculpins and misidentifications at the genus level. The remaining error levels for overall species-level identifications and identification of *Nocomis* species declined to levels well below the initial data quality objective of $<10\%$ measurement error. In this case, a re-analysis of preserved specimens and data records was used to characterize measurement error.

3.5. Guideline 9: Temporal variability — within the field season

For data collected from a large number of sites and documented as representing a single reporting period or season, it is essential to determine variability within the field season. For the benthic index, within-season temporal variability was characterized by revisiting sites in 13 estuaries during the same sampling season, and comparing the distribution of index values between the first and second visits (Fig. 6). If a site was found to be degraded (score < 3), marginal (score $=$ 3–5) or undegraded (score > 5) for both visits, then it would be plotted in either quadrant 2 or 4, or remain in the heavily shaded box (Fig. 6). If a site classification changed between visits, it would be plotted in quadrants 1 or 3, or in the lightly shaded zone. In this case, most sites retained their classification and only a few sites switched to or from a marginal classification. Correlation between the temporal replicates was significant ($P < 0.05$; $r = 0.83$). This validation, with additional information gathered from further statistical tests, showed that the benthic index was successful in demonstrating within-year temporal variability acceptable for meeting data quality objectives.

Establishing temporal variability was critical for the hypoxia indicator. Daily and short-term variations in bottom DO concentrations can vary dramatically at a single station. So this indicator, which uses a single point-in-time measurement, would be inappropriate for characterizing hypoxia at a specific station. Yet, the objective of the EMAP program was to evaluate ecological condition across a broad geographic scale, not at individual stations. Comparison of DO measurements taken at two different times during the index period showed the percent hypoxic area across the region to be nearly identical (Fig. 7). This stability assured that the indicator was appropriate for documenting spatial extent of hypoxia for large geographic regions.

3.6. Guideline 13: Data quality objectives

An indicator's discriminatory ability to meet data quality objectives, factoring in variability, precision and confidence levels desired by the program, should be quantified before incurring the costs associated with implementation in a monitoring program. For

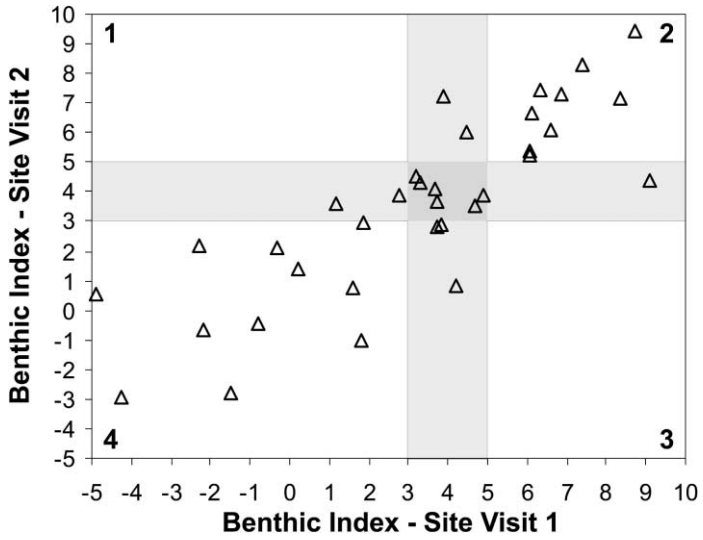

Fig. 6. Comparison of benthic index values from replicate site visits within a sampling season. Quadrant 2 indicates sites classified as undegraded for both visits; quadrant 4 indicates sites classified as degraded for both visits. Sites that fall within the gray shaded area (except for those sites in the center of the cross) changed from degraded or undegraded to marginal (or vice versa) from visits 1 to 2.

example, a program may require that an indicator be able to detect a 20% change in ecological condition over a 10-year period with 90% confidence. This determination is usually made by applying statistical power curves to data from a pilot study. In the *Evaluation Guidelines*, all three indicator examples employed statistical power curves as a means to evaluate the indicator's ability to meet the data quality objectives of the program. The program objective for the benthic index was an ability to detect a 2% trend (change) over 12 years with 90% confidence. Five power curves were computed (ranging from 1 to 3% trend magnitude) using data collected over a 4-year period (Fig. 8). According to these calculations, the performance goal was achievable using the benthic index; a 2% trend could be detected within approximately 10 years.

For the fish assemblage indicator, two performance objectives were considered. The indicator was required to detect trends in condition and distinguish classes of ecological condition within the proposed monitoring framework. A series of power curves was calculated to look at specific performance criteria (Fig. 9). From these curves it was determined that, after 4 years of monitoring, the standard error of the indicator score would range between one

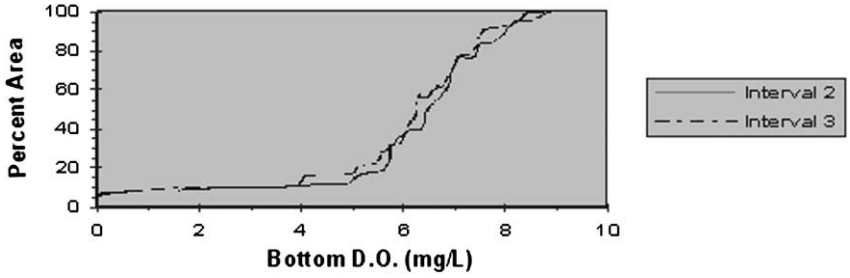

Fig. 7. A comparison of cumulative distribution functions for DO measurements made at the same sites at two different times (intervals 2 and 3) within the index period. This demonstrates relative stability of the hypoxia indicator across a large regional scale.

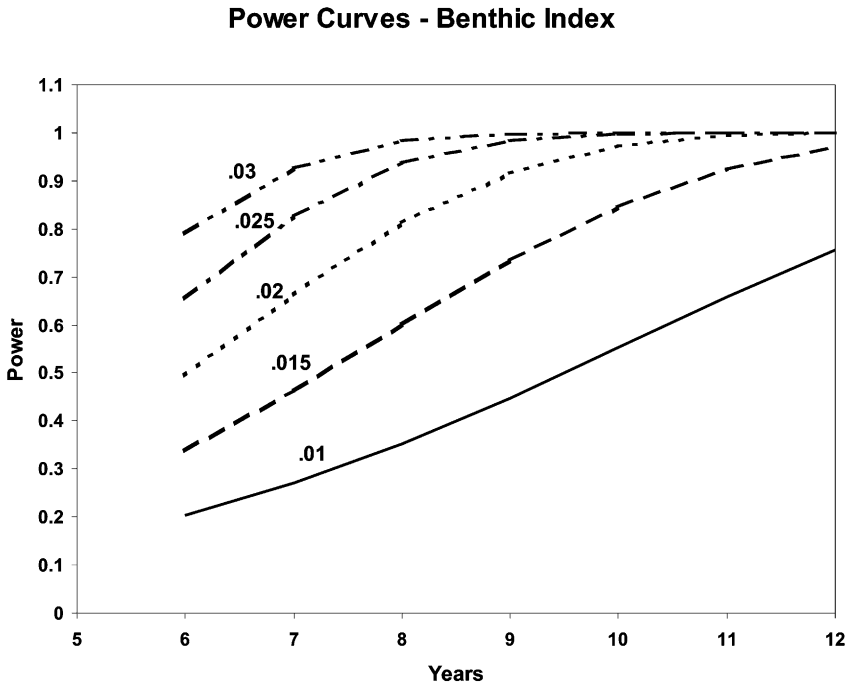

Fig. 8. Power curves for detecting temporal variability across years of 1–3% in the proportion of area with degraded benthic communities relevant to the benthic index example.

and two points. These results implied that three or four classes of condition could be distinguished over the potential range of indicator scores, thereby satisfying the performance criterion (Fig. 9A). Further analysis demonstrated that magnitude of coherent across-year variance (Fig. 9B) and the desired magnitude of change (Fig. 9C) had more effect on the ability of the indicator to detect trend than did sample size (Fig. 9D). Status and trend detection was considered possible with this indicator if across-year variance was relatively small compared to within-year variance.

3.7. Guideline 14: Assessment thresholds

Establishing thresholds for management action is an essential role for indicators. For some, such as the hypoxia indicator, thresholds may be pre-determined by regulation. Several states have adopted 2 mg/l (point-in-time) and 5 mg/l (24 h continuous) minimum thresholds for DO concentrations in estuaries, levels shown to be necessary for a healthy ecosystem. For the benthic index, thresholds were established by comparing numeric index values for each site with the original a priori classifications, degraded or undegraded, which were based on measurements of sediment contaminants, sediment toxicity and hypoxia. A cumulative distribution curve (Fig. 10) exposed overlap between a priori designations and benthic index values, so a 'marginal' zone was created for index values between 3 and 5. For the fish assemblage indicator, thresholds were proposed for individual metrics as well as for the final indicator score. Thresholds for individual metrics were derived either from values obtained at reference sites or from the distribution of values obtained from a large regional study (the 1993–1994 Mid-Atlantic Highlands Assessment) that included numerous sample sites. Distribution of values from the same study were used to define four 'operational' classifications (excellent, acceptable, marginal and impaired) for the final indicator score; these classifications and thresholds must still be quantitatively validated.

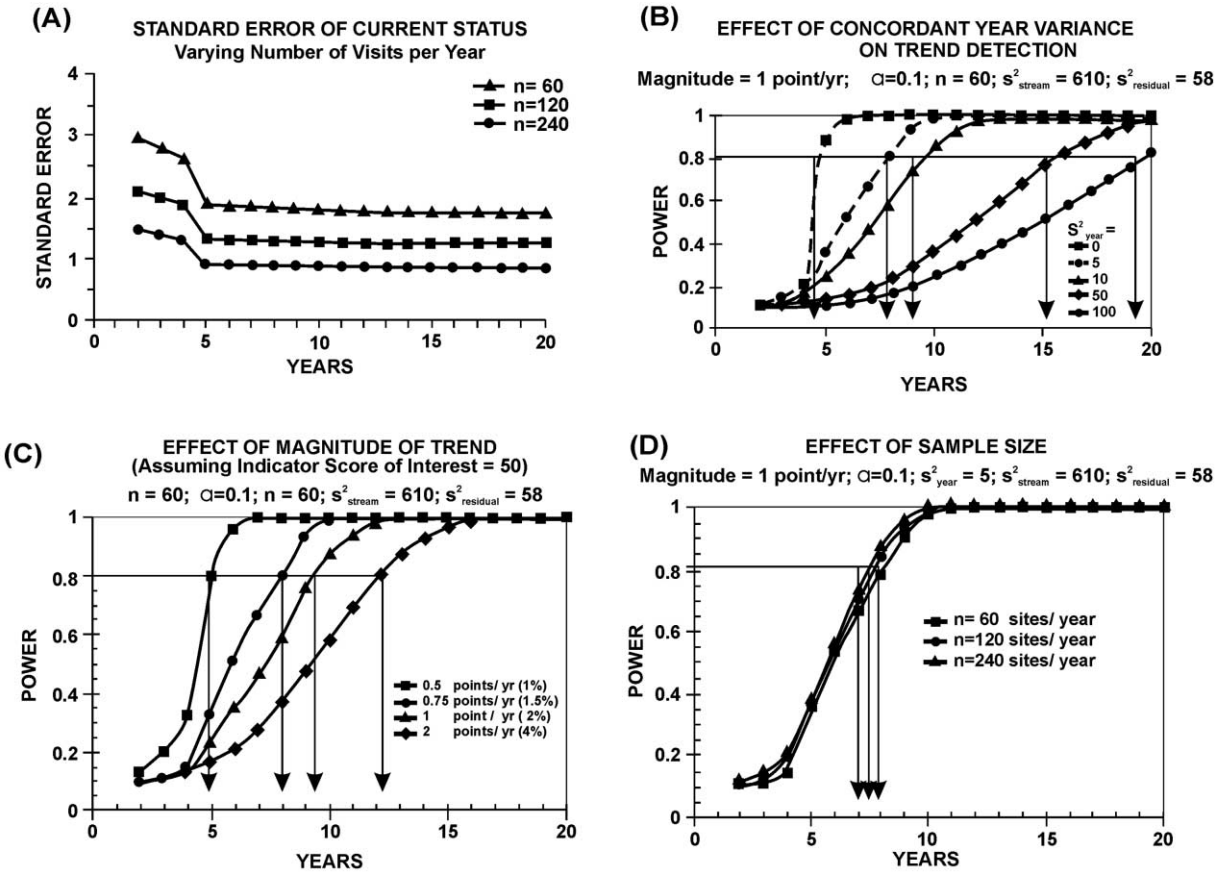

Fig. 9. Statistical power curves for the fish assemblage indicator. (A) Effect of annual sample size on standard estimate of indicator score. (B) Effect of the magnitude of coherent across-year variance (indicator units) on trend detection. (C) Capability to detect different magnitudes of trend. (D) Effect of annual sample size on trend detection.

3.8. Guideline 15: Linkage to management action

Indicators are useful only if they can provide information to support a management decision or to quantify the success of past decisions. Policy makers and resource managers must be able to recognize the implications of indicator results for stewardship, regulation or research. This requires that an indicator be reasonably understandable to the public, have a link to policy measures or display some utility in cost–benefit assessments. The hypoxia indicator is a direct measure of the geographic extent of the stressor. As such, it provides a direct link to management actions attempting to control nutrient effluent from sewage treatment plants, septic tanks, and non-point sources.

The benthic index was identified as useful to environmental managers and policy and decision makers who want to identify areas of potential degradation, or track the status of environmental condition over time. This indicator provides a quantification of the response of benthic communities to environmental stress (Summers et al., 1995). It is intended to provide assessments of estuarine condition across large geographic areas. Since the benthic index is scalable and the criteria for determining a site's classification (degraded or undegraded) are numeric, application of the index to various estuaries is straightforward. It can be used to answer questions about the status of benthic communities in the estuaries of large geographic regions, the spatial or temporal variation of degraded areas of benthic communities, and the ecological condition

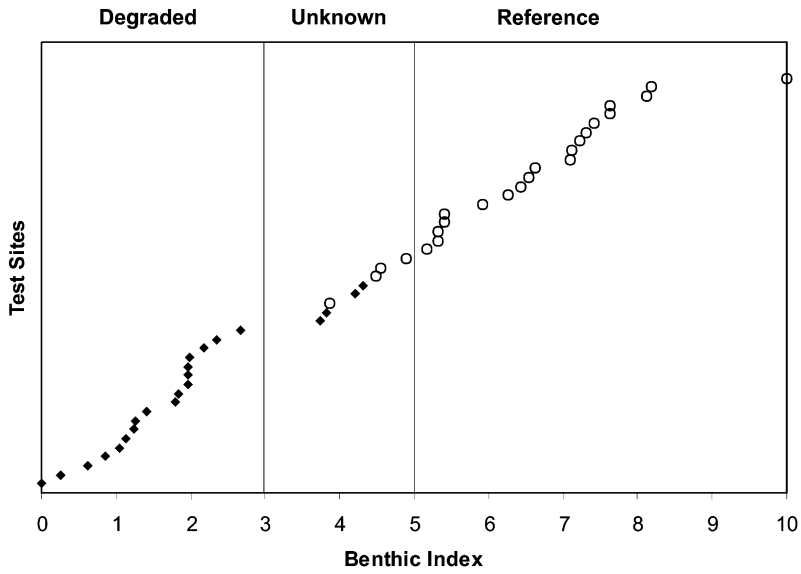

Fig. 10. Distribution of benthic index values for test sites ((◆) degraded sites; (○) undegraded sites), determined by a priori criteria for dissolved oxygen, sediment chemistry, and sediment toxicity. The area of overlap, between index values of 3 and 5, was designated 'marginal'.

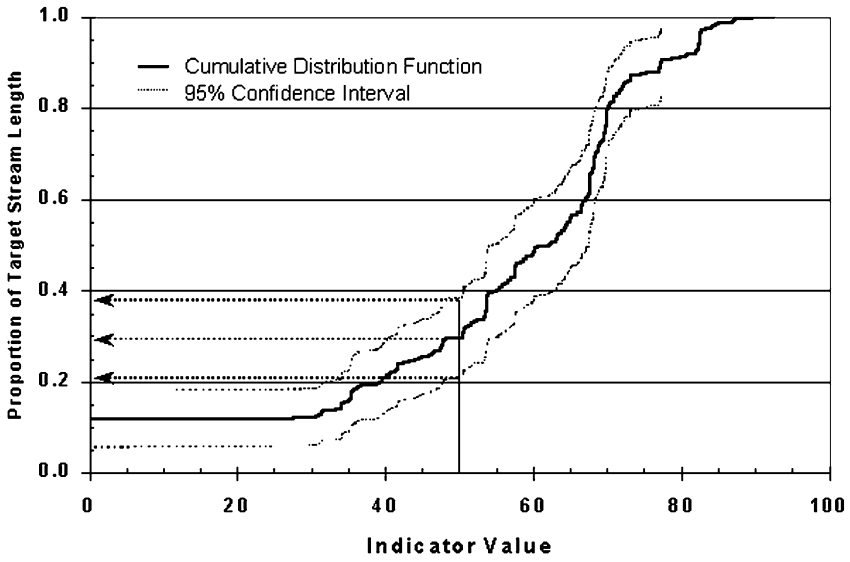

Fig. 11. A hypothetical example of how the fish assemblage indicator might be used by resource managers to estimate the status of the resource population.

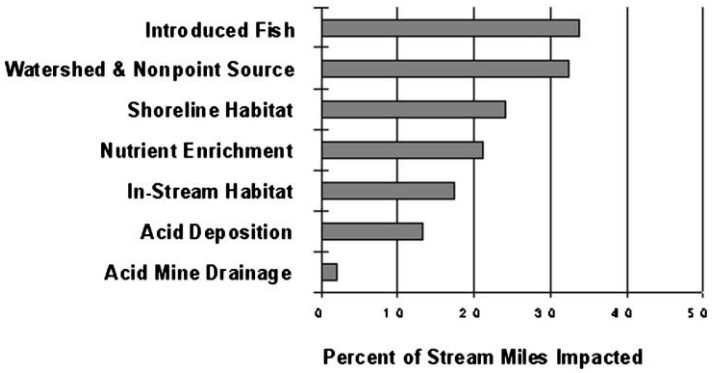

Fig. 12. Relative ranking of stressor variables identified for the fish assemblage indicator, based on proportion of target resource population impacted.

of benthic communities in estuaries of different regions.

The fish assemblage example also demonstrated the cumulative distribution of indicator values as a means of communicating with managers and policy makers (Fig. 11). Because of the multiple components in the indicator, it was also possible to delineate and rank the influences of various stressors (Fig. 12). It was found, using data collected during 1993 and 1994, that introduced fish species and watershed-level disturbances were the most regionally extensive stressors in the region studied, whereas acidic deposition, although often considered a greater threat where it occurs, was less extensive in the region.

4. Discussion

The *Evaluation Guidelines* (US EPA, 2000; Table 1) offers a consistent means to highlight strengths and weaknesses of indicators within the context of specific program objectives. This is important for comparison and selection among potential indicators and for the iterative improvement of indicators during development. However, the example indicators employed in the document also demonstrate flexibility in the approach and type of information used to address the guidelines. This robust quality is underscored here because it implies that the *Evaluation Guidelines* may be useful to a variety of programs. The guidelines are not criteria and do not, by themselves, determine indicator applicability or effectiveness. Rather, they provide the framework for asking relevant questions about an indicator. Users decide whether or not the indicator is acceptable based on the objectives and resources of their specific programs.

The examples varied from a simple chemical measurement to a complex multimetric index, yet the *Evaluation Guidelines* were applied to all three with reasonable success. The evaluation process identified temporal and spatial variation in DO measurements as critical for determining areal extent of hypoxia; in particular, spatial expansion improved indicator stability. The benthic index was constructed from statistical evaluations of environmental data, so achievement of data quality objectives was anticipated. However, more research is needed to determine assessment thresholds (Guideline 14) since site classifications are currently based on a priori criteria for dissolved oxygen, sediment chemistry and sediment toxicity. In the fish assemblage example, it was found that responsiveness of certain components in the indicator did not meet the data quality objectives. Since some of these are believed to have inherent biological value, a decision must be made to accept, exclude or improve the response of those components.

Investigation and protection of ecological resources continues to change in focus and complexity. In keeping with these changes, a dynamic battery of useful and efficient indicators is essential. The *Evaluation Guidelines*, which provides a consistent and robust framework for indicator evaluation, can become an important tool in the continuous process of developing ecological indicators.

Acknowledgements

The EPA report, *Evaluation Guidelines for Ecological Indicators*, is available free of charge by calling +1-800-490-9198, and requesting document #EPA/620/R-99/005. The electronic version may be downloaded from the EMAP website at www.epa.gov/emap/. We appreciate the contribution of the hypoxia indicator example provided by Charles J. Strobel, US EPA, and James Heltsche, OAO Corporation, Narragansett, RI; the benthic index example provided by Virginia D. Engle, US EPA, and the fish assemblage indicator provided by Frank H. McCormick, US EPA, and David V. Peck, US EPA. This article is Gulf Ecology Division Contribution no. 1133 of the US Environmental Protection Agency, Gulf Ecology Division (GED), Office of Research and Development, National Health and Environmental Effects Laboratory (NHEERL), Gulf Breeze, FL.

References

Bardwell, L.V., 1991. Problem-framing: a perspective on environmental problem-solving. Environmental Management 15, 603–612.

Cowling, E.B., 1992. The performance and legacy of NAPAP. Ecological Applications 2, 111–116.

Karr, J.R., Fausch, K.D., Angermeier, P.L., Yant, P.R., Schlosser, I.J., 1986. Assessing Biological Integrity in Running Waters: A Method and its Rationale. Illinois Natural History Survey Special Publication No. 5.

McCormick, F.H., Hughes, R.M., 1998. Aquatic vertebrate indicator. In: Klemm, D.J., Lazorchak, J.M., Peck, D.V. (Eds.), Environmental Monitoring and Assessment Program — Surface Waters: Field Operations and Methods for Measuring the Ecological Condition of Wadeable Streams. EPA/620/R–94/004, US Environmental Protection Agency, Cincinnati, OH.

Summers, J.K., Paul, J.F., Robertson, A., 1995. Monitoring the ecological condition of estuaries in the United States. Toxicological and Environmental Chemistry 49, 93–108.

Thornton, K.W., Saul, G.E., Hyatt, D.E., 1994. Environmental Monitoring and Assessment Program: Assessment Framework. EPA/620/R-94/016, US Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC.

United States Environmental Protection Agency (US EPA), 1994. In: Barber, M.C. (Ed.), Environmental Monitoring and Assessment Program Indicator Development Strategy. EPA/620/R-94/022, US Environmental Protection Agency, Office of Research and Development, Washington, DC.

United States Environmental Protection Agency (US EPA), 2000. In: Jackson, L.E., Kurtz, J.C., Fisher, W.S. (Eds.), Evaluation Guidelines for Ecological Indicators. EPA/620/R-99/005, US Environmental Protection Agency, Office of Research and Development, Research Triangle Park, NC.